OD SUROWYCH DANYCH DO PROFESJONALNEGO RAPORTU – PRZYKŁAD KOMPLETNEJ STATYSTYCZNEJ ANALIZY DANYCH MEDYCZNYCH

Janusz Wątroba, StatSoft Polska Sp. z o.o.

Punktem wyjścia przedstawianego referatu jest próba przeciwstawienia się dość powszechnie występującym poglądom na temat znaczenia i rzeczywistych korzyści, jakie przynosi stosowanie metod statystycznej analizy danych. W tym celu analiza danych zostanie osadzona w szerszym kontekście procesu badawczego w dziedzinach badań wykorzystujących empiryczne dane. Szczególny nacisk położono na zaprezentowanie typowych etapów analizy danych jako ciągu wzajemnie powiązanych logicznych operacji prowadzących od surowych danych do końcowych wyników stanowiących podstawę do wyciągania merytorycznych wniosków ściśle związanych z celem badań. Całość zostanie zilustrowana opracowaniem przykładowych danych pochodzących z badań medycznych. Przy opracowywaniu danych wykorzystano niektóre z licznych narzędzi i rozwiązań, które do wspomagania tego procesu oferuje środowisko programu *STATISTICA*.

Wstęp

Wśród opinii wyrażanych na temat analizy danych (zwłaszcza statystycznej) można napotkać wiele stereotypów. Najczęstsze z nich to: (1) przekonanie, że analizę danych można poprawnie przeprowadzić w oderwaniu od całości procesu badawczego, towarzyszącego realizacji określonych celów oraz bez pewnej ogólnej orientacji w dziedzinie, której badania dotyczą, i znajomości postawionych przez badacza szczegółowych pytań lub hipotez, (2) traktowanie analizy danych jak „czarnej skrzynki", do której należy tylko włożyć niemalże dowolne dane, a na wyjściu otrzymuje się gotowe do merytorycznej interpretacji wyniki i (3) poznawanie technik analizy danych i samodzielne ich stosowanie jest niepotrzebną stratą czasu i wysiłku. Konsekwencją takiej sytuacji jest z jednej strony dość powszechne traktowanie analizy danych jako swoistego kwiatka do butonierki, a z drugiej jako zbioru technik, za pomocą których można niemalże wszystko udowodnić i to w dodatku w oparciu o empirycznie zgromadzone dane. Powiedzenie „kłamstwo, duże kłamstwo, statystyka" jest tego najlepszym potwierdzeniem.

Oprócz przytoczonych wyżej opinii można napotkać również inne. Wśród wielu osób zainteresowanych analizą danych, ale pracujących poza instytucjami naukowymi, panuje również przekonanie, że zwłaszcza statystyczna analiza danych może dać określone

Copyright © StatSoft Polska 2006

korzyści praktyczne, ale tylko w przypadku badań naukowych, gdzie zazwyczaj chodzi głównie o zrealizowanie pewnych celów poznawczych.

Znaczenie analizy danych w badaniach empirycznych

Zacznijmy od przeciwstawienia się pierwszemu z podanych wyżej stereotypów na temat analizy danych. Spróbujmy uzasadnić, że analiza danych nie powinna być przeprowadzana w oderwaniu od całości procesu, który możemy określić jako badania empiryczne. Analiza danych jest jednym z ważnych etapów procesu badawczego. Umieszczenie jej w szerszym kontekście (procesu badawczego) ma na celu zwrócenie uwagi na to, że etapy, które ją poprzedzają, mają olbrzymi wpływ na jej przebieg. Niektóre informacje o wcześniejszych etapach procesu badawczego są wręcz konieczne do tego, aby poprawnie wybrać metody analizy. Chodzi na przykład o takie kwestie jak to, które ze zmiennych traktujemy jako niezależne (objaśniające, przyczyny), a które jako zależne (objaśniane, skutki). Kolejna rzecz to charakter zmiennych. Tutaj chodzi o to, które ze zmiennych mają charakter jakościowy, a które ilościowy. Następne pytanie dotyczy tego, czy analizowane jednostki (obiekty) stanowią kompletny zbiór (populację) czy też stanowią pewną reprezentację tego zbioru. Jeśli tak to kolejne pytanie brzmi: czy badany zbiór jednostek jest reprezentatywny dla całej populacji, dla której chcemy prowadzić wnioskowanie czy też stanowi tylko pewien nieokreślony jej podzbiór (czyli występuje brak informacji o reprezentatywności badanego zbioru). Z drugiej strony nawet poprawne przeprowadzenie analizy nie powinno kończyć się na suchym podaniu wyników, ponieważ wyniki analizy danych nie przełożone na merytoryczne wnioski (odpowiadające postawionym na wstępnie lub później hipotezom i pytaniom badawczym) są dla badacza bezużyteczne.

Planowanie i realizacja całego procesu badawczego podlega dość rygorystycznym zasadom opisywanym w wielu podręcznikach poświęconych metodologii badań empirycznych. Najogólniej można chyba stwierdzić, że głównym celem procesu badawczego jest zazwyczaj *zdobycie (lub potwierdzenie) wiedzy* na temat określonego wycinka otaczającej nas rzeczywistości przyrodniczej lub społecznej. Proces badawczy najczęściej jest przedstawiany w postaci cyklu działań. Analiza danych jest traktowana jako jedno z tych działań.

Typowy przebieg analizy danych

Próba stworzenia pewnego typowego schematu analizy danych napotyka na spore problemy. Powodem jest to, że każda konkretna analiza ma swoją specyfikę. Wydaje się jednak, że można wskazać na pewne typowe działania, które są wykonywane w przypadku każdej analizy danych. Przystępując do przeprowadzenia analizy, zawsze trzeba zacząć od *przygotowania danych do analizy*. Etap ten obejmuje zazwyczaj takie czynności jak: utworzenie elektronicznej wersji zbioru danych, scalenie danych pochodzących z różnych źródeł, sprawdzenie i ujednolicenie sposobu kodowania braków danych oraz sposobu zapisu wartości zmiennych jakościowych oraz sprawdzenie danych pod kątem wartości nietypowych.

 Copyright © StatSoft Polska 2006

Kolejny etap to *dobór odpowiednich metod opracowania danych*. Jest to niewątpliwe etap, który zazwyczaj sprawia najwięcej trudności, gdyż z jednej strony wymaga on dobrej merytorycznej znajomości badanych zjawisk i zgromadzonych danych empirycznych, a z drugiej strony sporego doświadczenia w zakresie różnych metod statystycznej analizy danych. Oprócz tego zdefiniowanie analizy wymagać może nieco dokładniejszej wiedzy na temat parametrów analizy. Oczywiście nieodzowne są także dostęp i znajomość obsługi odpowiednich narzędzi informatycznych. Poprawne zdefiniowanie i przeprowadzenie analizy nie oznacza, że proces jest zakończony.

Ostatni etap polega na *przygotowaniu wyników analizy do ich merytorycznej interpretacji* pod kątem postawionych pytań i hipotez badawczych. Wymaga to przede wszystkim zestawienia najważniejszych wyników analizy w postaci raportu, ale też praktycznej pomocy badaczowi w przeprowadzeniu interpretacji wyników analiz statystycznych.

Przedstawione uwagi na temat etapów analizy danych zostaną zilustrowane praktycznym przykładem w środowisku programu *STATISTICA*.

Przykład kompletnej analizy danych w programie *STATISTICA*

Przedmiotem analizy będą dane dotyczące wybranych parametrów biochemicznych i klinicznych zebranych dla pacjentów, u których zdiagnozowano występowanie choroby niedokrwiennej serca. Wśród części badanych osób wystąpił zawał mięśnia sercowego. Ponadto dane zawierają informację na temat rodzaju przebytego zawału (niepełnościenny lub pełnościenny).

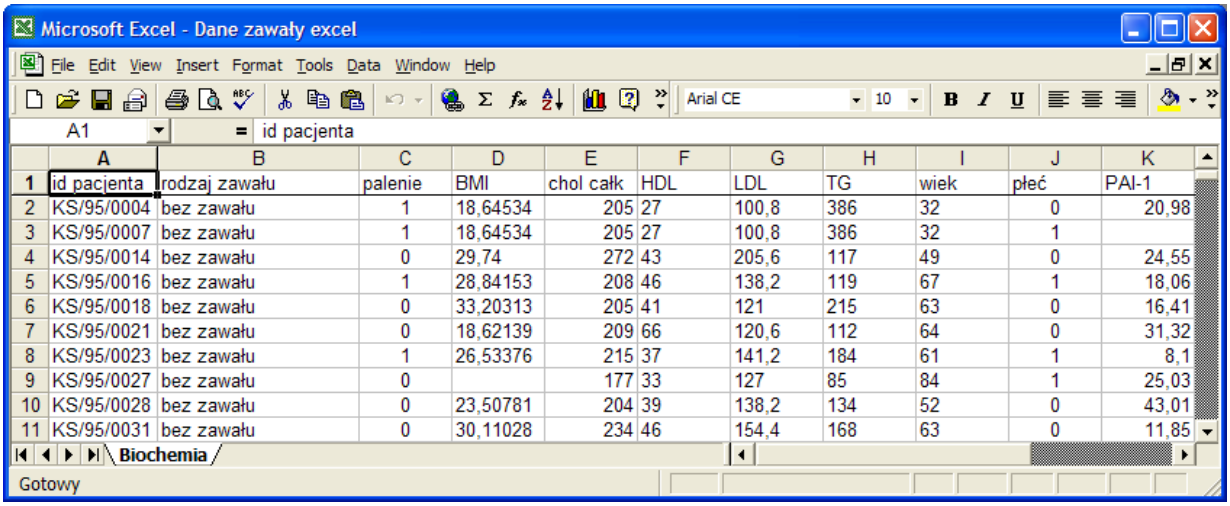
Główne cele badawcze prezentowanej analizy to:

- ocena wpływu płci i palenia na ryzyko wystąpienia zawału i jego rodzaj,
- ocena niezależnego wpływu badanych parametrów biochemicznych i klinicznych na ryzyko i rodzaj zawału oraz
- ocena łącznego wpływu branych pod uwagę czynników na ryzyko i rodzaj zawału.

Opracowanie zebranych danych rozpoczniemy od etapu *przygotowania danych do analizy*. W rzeczywistych badaniach bardzo często zdarza się sytuacja, w której dane są zgromadzone w kilku zbiorach przygotowanych za pomocą różnych aplikacji (najczęściej w postaci pliku Excela, bazy danych lub pliku tekstowego). Podobnie jest w prezentowanym przykładzie. Część danych została zgromadzona w arkuszu programu Excel. Fragment danych przedstawiono na rysunku poniżej.

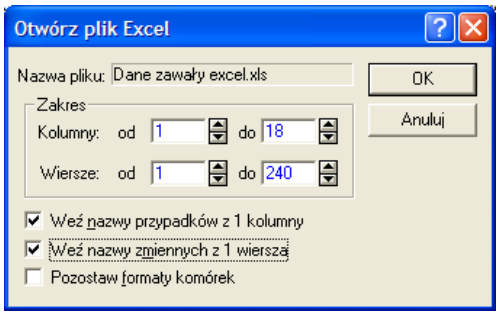
W związku z tym, przygotowując dane do dalszego opracowania, rozpoczynamy od zaimportowania zbioru (lub zbiorów) danych. Do poprawnego przeprowadzenia tej operacji wystarczają zazwyczaj podstawowe informacje na temat struktury danych (np. czy zbiór zawiera jakieś identyfikatory badanych obiektów oraz nazwy badanych parametrów).

Copyright © StatSoft Polska 2006

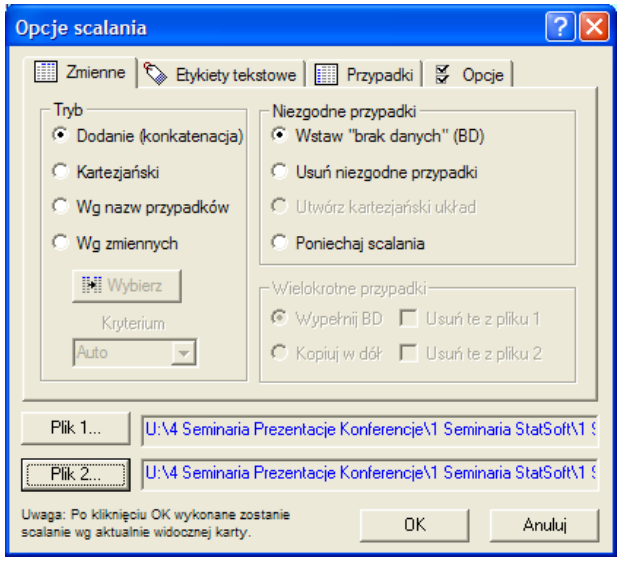


	A	B	C	D	E	F	G	H	I	J	K
1	id pacjenta	rodzaj zawału	palenie	BMI	chol całk	HDL	LDL	TG	wiek	płeć	PAI-1
2	KS/95/0004	bez zawału	1	18,64534	205	27	100,8	386	32	0	20,98
3	KS/95/0007	bez zawału	1	18,64534	205	27	100,8	386	32	1	
4	KS/95/0014	bez zawału	0	29,74	272	43	205,6	117	49	0	24,55
5	KS/95/0016	bez zawału	1	28,84153	208	46	138,2	119	67	1	18,06
6	KS/95/0018	bez zawału	0	33,20313	205	41	121	215	63	0	16,41
7	KS/95/0021	bez zawału	0	18,62139	209	66	120,6	112	64	0	31,32
8	KS/95/0023	bez zawału	1	26,53376	215	37	141,2	184	61	1	8,1
9	KS/95/0027	bez zawału	0		177	33	127	85	84	1	25,03
10	KS/95/0028	bez zawału	0	23,50781	204	39	138,2	134	52	0	43,01
11	KS/95/0031	bez zawału	0	30,11028	234	46	154,4	168	63	0	11,85

Poniżej pokazano typowy wygląd okna pojawiającego się w trakcie importu arkusza programu Excel do programu *STATISTICA*.



Gdy oryginalne dane występują w kilku osobnych zbiorach, wówczas po ich zaimportowaniu zachodzi potrzeba ich scalenia w jeden plik danych. W programie *STATISTICA* mamy do dyspozycji wiele użytecznych, a jednocześnie łatwych do wykorzystania opcji pozwalających na scalanie kilku zbiorów danych według zmiennych, przypadków lub etykiet tekstowych. Poniżej zamieszczono okno, które pojawia się przy scalaniu arkuszy.



 Copyright © StatSoft Polska 2006

Efektem tych operacji ma być jeden zbiór danych, który zawiera wszystkie informacje o zgromadzonych danych, czyli wszystkie badane przypadki i zmienne, dla których przeprowadzono pomiary lub obserwacje. Dzięki temu można przeprowadzać różne operacje i analizy na danych bez potrzeby sięgania do innych zbiorów. Poniżej zamieszczono widok fragmentu arkusza danych zawierającego zaimportowane i scalone dane z opisywanego przykładu.

Dane: DaneZawały (17 zmn. * 239 prz.)

Dane dotyczą wybranych parametrów biochemicznych oraz klinicznych zebranych dla pacjentów z chorobą niedokrwienną serca.
Dane pochodzą z książki prof. Watały Biostatystyka - wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych, α-medica press 2002.

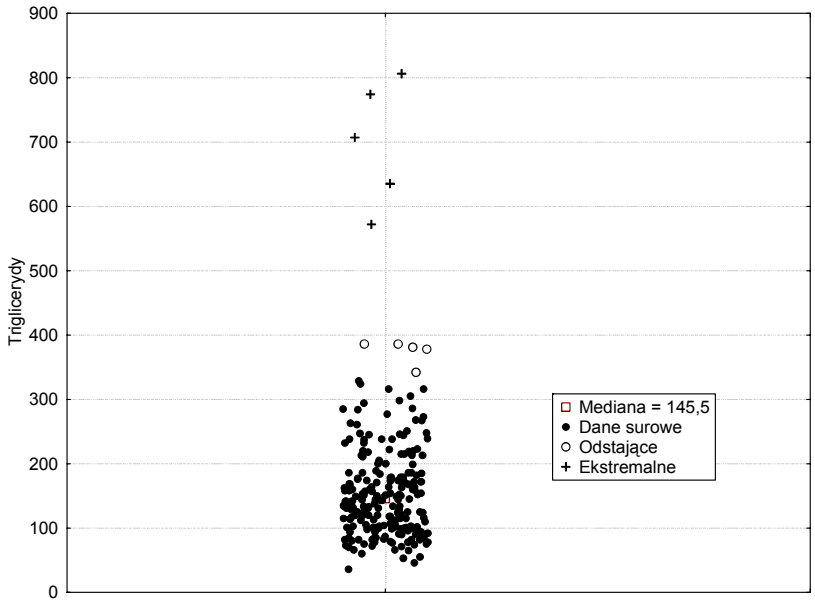
	1 Płeć	2 Wiek	3 Choroba wiencowa	4 Czas choroby wiencowej	5 Rodzaj zawału	6 Palenie	7 BMI	8 Cholesterol całkowity	9 HDL	10 LDL
KS/95/0004	kobieta	32	tak	poniżej 2 m-cy	bez zawału	pali	18,65	205	27	100,8
KS/95/0007	mężczyzna	32	tak	poniżej 2 m-cy	bez zawału	pali	18,65	205	27	100,8
KS/95/0014	kobieta	49	tak	od 2 do 12 m-cy	bez zawału	nie pali	29,74	272	43	205,6
KS/95/0016	mężczyzna	67	tak	od 2 do 12 m-cy	bez zawału	pali	28,84	208	46	138,2
KS/95/0018	kobieta	63	tak	powyżej 12 m-cy	bez zawału	nie pali	33,20	205	41	121,0
KS/95/0021	kobieta	64	tak	poniżej 2 m-cy	bez zawału	nie pali	18,62	209	66	120,6
KS/95/0023	mężczyzna	61	tak	poniżej 2 m-cy	bez zawału	pali	26,53	215	37	141,2
KS/95/0027	mężczyzna	84	nie		bez zawału	nie pali		177	33	127,0
KS/95/0028	kobieta	52	tak	powyżej 12 m-cy	bez zawału	nie pali	23,51	204	39	138,2

W następnej kolejności połączony zbiór danych może wymagać sprawdzenia i ewentualnego ujednolicenia sposobu kodowania brakujących danych oraz sposobu kodowania zmiennych jakościowych. Do tego celu w programie *STATISTICA* są dostępne odpowiednie narzędzia, m. in. Edytor etykiet tekstowych oraz opcja Przekodowania.

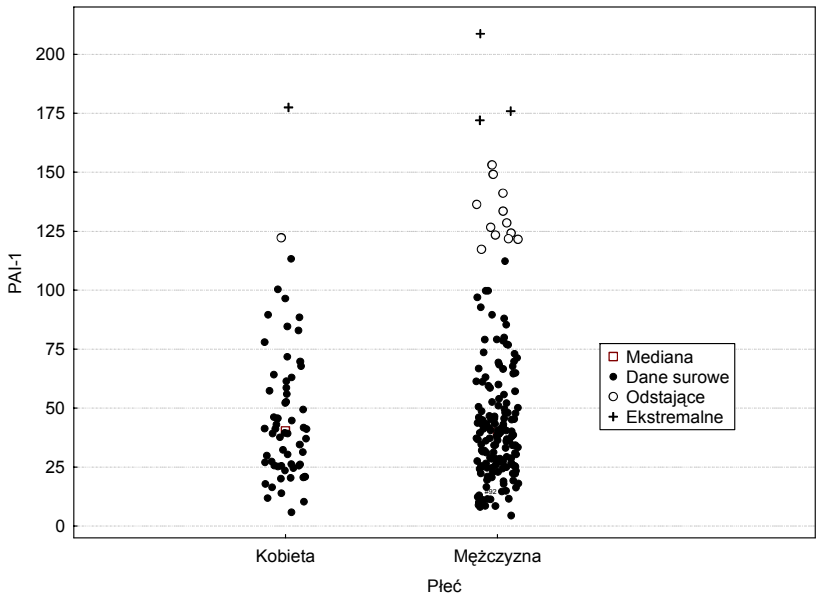
Kolejną bardzo ważną, bo mającą wpływ na wyniki analizy i ich merytoryczną interpretację rzeczą jest sprawdzenie danych pod kątem nietypowych (odstających) obserwacji. Trzeba w tym miejscu wyraźnie zaznaczyć, że nie chodzi tu wyłącznie o znalezienie ewidentnych błędów wartości pomiarów lub obserwacji (np. wiek osoby - 200 lat), lecz także o wskazanie obserwacji odbiegających od pozostałych, które choć formalnie poprawne, mogą jednak wypaczać wyniki analizy. W związku z tym warto na tym etapie poświęcić trochę czasu, aby uniknąć problemów w przyszłości. Jednocześnie sprawdzanie danych pod kątem poprawności zmusza do przeglądania danych, co może zaowocować lepszym „zrozumieniem" danych i rozwiązywanego problemu badawczego (bardzo często na tym etapie badacze zauważają i zaznaczają konkretne obserwacje, które pozwalają na lepsze poznanie złożoności badanego zjawiska). Program *STATISTICA* oferuje w tym względzie wygodne i bardzo efektywne narzędzia graficzne i analityczne. Metody analityczne są stosowane zazwyczaj wtedy, gdy badacz ma dobrą wiedzę na temat tego, jakie zakresy wartości przyjmują badane zmienne. W przypadku danych medycznych bardzo przydaje się znajomość norm. Jeśli badacz nie posiada wystarczającej wiedzy na ten temat, wówczas może zastosować do sprawdzania danych różne graficzne techniki eksploracyjne, zaprojektowane do badania danych pod kątem odstających obserwacji.

Najbardziej popularne graficzne techniki badania danych to wykresy typu ramka-wąsy oraz wykresy rozrzutu. Zamieszczone poniżej rysunki pokazują przykładowe wykresy utworzone dla danych z opisywanego przykładu.

Copyright © StatSoft Polska 2006



Powyższy wykres ilustruje rozkład wartości poszczególnych obserwacji zmiennej Triglicerydy. Puste kółeczka oznaczają obserwacje odstające, tzn. takie, których położenie odbiega od środka rozkładu. Punkty oznaczone znakami plus oznaczają z kolei tzw. obserwacje ekstremalne, które zdecydowanie odbiegają od środka rozkładu. Oczywiście uznanie danej obserwacji za nietypową jest zawsze arbitralne, dlatego badacz za każdym razem musi podejmować samodzielną decyzję o tym, jak będzie traktował konkretną obserwację. Wykres pozwala tylko te obserwacje zidentyfikować. Rysunek pokazuje, że w analizowanym zbiorze danych wystąpiło 5 obserwacji zdecydowanie odstających od środka rozkładu. Wykresy tego typu można także w łatwy sposób tworzyć dla określonych grup przypadków.



 Copyright © StatSoft Polska 2006

Powyżej pokazano przykład takiego wykresu, zawierającego rozkład wartości zmiennej PAI-1 z uwzględnieniem płci badanych pacjentów. Jak widać, wśród badanych mężczyzn występuje nieznaczna tendencja do występowania nieco większej liczby nietypowych obserwacji. Nasuwa się pytanie, czy jest to jakaś prawidłowość czy też stało się tak przypadkowo (oczywiście przy założeniu, że traktujemy wartości jako poprawne).

Na etapie sprawdzania danych może zachodzić potrzeba sięgania do danych źródłowych i kontaktu z osobami, które były odpowiedzialne za ich gromadzenie. Warto również wiedzieć, że po stwierdzeniu, że dana obserwacja jest nietypowa, ale poprawna, możemy w arkuszu danych umieścić odpowiedni zapis dla tej konkretnej obserwacji, który jest przechowywany razem z arkuszem.

Dopiero po wielostronnym sprawdzeniu danych powinno się doliczać ewentualne nowe zmienne, np. jakieś wskaźniki lub indeksy obliczane na podstawie wcześniej wprowadzonych zmiennych. Chodzi o to, żeby uniknąć ewentualnego przenoszenia błędów (sprawdzenie dodatkowych zmiennych pod kątem odstających obserwacji też na pewno nie zaszkodzi analizie).

Wydaje się, że kluczowym, a jednocześnie sprawiającym najwięcej trudności etapem analizy danych jest *dobór odpowiednich metod opracowania danych*. Aby go poprawnie przeprowadzić, należy brać pod uwagę kilka kwestii. Na poczatku tego etapu wskazane jest zrekonstruowanie całego badania lub jego zasadniczych fragmentów. Trzeba zestawić wszystkie dostępne w pliku danych zmienne, określić ich charakter (jakościowe, ilościowe), ustalić relacje pomiędzy badanym zbiorem obiektów a populacją i skonfrontować te informacje z postawionymi merytorycznymi celami badań. Zasadnicza kwestia to określenie celu badań w języku technik analizy danych. Wymaga to od badacza uszczegółowienia a czasem nawet przeformułowania niektórych pytań lub hipotez w świetle zgromadzonych danych. Dla ilustracji poniżej podano cele badawcze z opisywanego przykładu przełożone na język analizy danych:

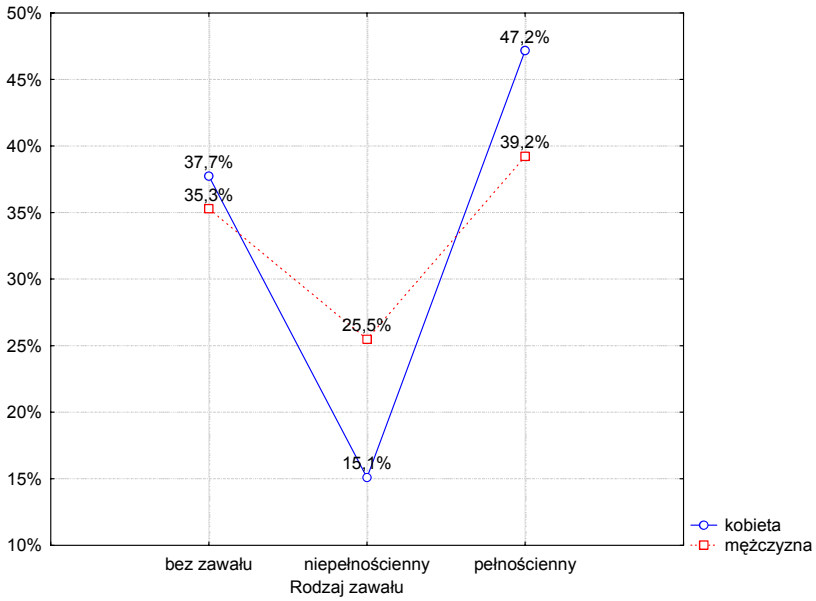
- ocena zróżnicowania częstości występowania zawału i jego rodzajów w zależności od płci badanych osób i statusu palenia,
- ocena zróżnicowania przeciętnego poziomu badanych parametrów biochemicznych i klinicznych w grupach pacjentów, u których wystąpił bądź nie wystąpił zawał, oraz w grupach pacjentów, u których stwierdzono różne rodzaje zawałów oraz
- budowa całościowego modelu opisującego łączny wpływ parametrów biochemicznych i klinicznych na ryzyko wystąpienia i rodzaj zawału.

Przykładowo dla zrealizowania pierwszego z przedstawionych celów należy utworzyć tabele dwudzielcze dla odpowiednich zmiennych. Poniżej zamieszczono taką tabelę dla zmiennych Płeć i Rodzaj zawału.

Copyright © StatSoft Polska 2006

Płeć	Podsumowująca tabela dwudzielcza			
	Rodzaj zawału bez zawału	**Rodzaj zawału niepełnościenny**	Rodzaj zawału pełnościenny	Wiersz Razem
kobieta	20	8	25	53
%Wiersza	37,74%	15,09%	47,17%	
mężczyzna	54	39	60	153
%Wiersza	35,29%	25,49%	39,22%	
Ogół	74	47	85	206

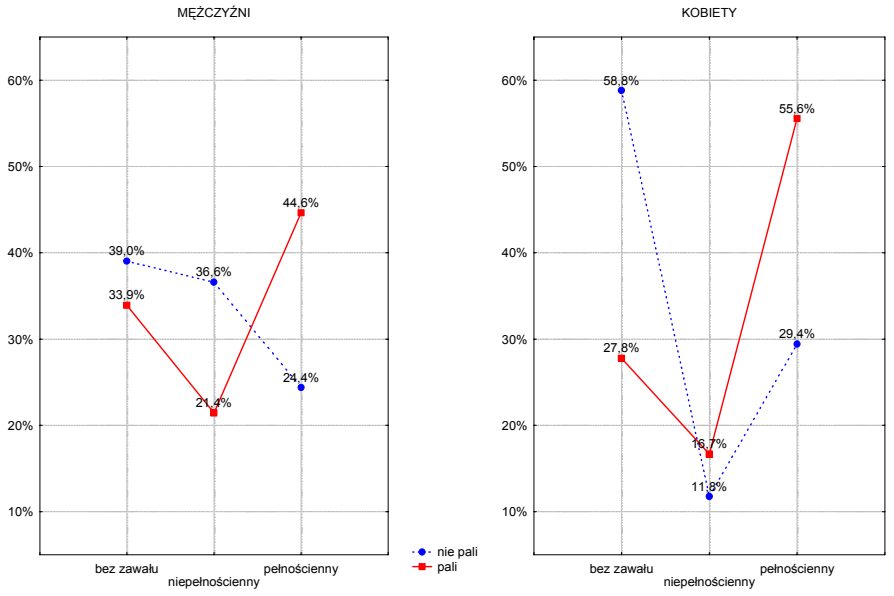
Tabela pokazuje, że liczebności pacjentów ze względu na rodzaj zawału nie były równoliczne (są one umieszczone w ostatnim wierszu tabeli z wynikami). Jednocześnie przedstawia ona odsetki pacjentów, u których stwierdzono różne rodzaje zawału (lub jego niewystąpienie) osobno w odniesieniu do ogólnej liczby badanych kobiet i mężczyzn. Dzięki temu możemy stwierdzić między innymi, że u badanych mężczyzn częściej notowano przypadki wystąpienia zawału pełnościennego (o około 8%) niż u badanych kobiet. Poniżej dla graficznej ilustracji zamieszczono odpowiedni wykres.



Podobnie można ocenić wpływ palenia. Jednak dużo ciekawsze wydaje się jednoczesne wzięcie pod uwagę płci i statusu palenia i ocena ich jednoczesnego wpływu na rodzaj zawału. Poniżej zamieszczono wykres, który pokazuje otrzymane wyniki takiej analizy.

Przedstawione wyniki pokazują między innymi, że u palących kobiet częściej obserwowano zawał pełnościenny niż u palących mężczyzn (różnica wynosi ponad 10%). Czy można traktować ten wynik jako dowód na występowanie określonej prawidłowości?

 Copyright © StatSoft Polska 2006



Dla realizacji drugiego z celów analizy należy obliczyć podstawowe statystyki opisowe dla branych pod uwagę zmiennych ilościowych w grupach według danej zmiennej jakościowej (np. Rodzaj zawału). Analiza tego typu jest określana terminem „analiza przekrojowa". Poniżej zamieszczono tabelę z przykładowymi wynikami dla zmiennej PAI-1.

Tabela z wynikami analizy przekrojowej

Rodzaj zawału	PAI-1 N	PAI-1 Średnia	**PAI-1 Odch.std.**	PAI-1 Mediana	PAI-1 Q25	PAI-1 Q75
bez zawału	72	33,53	21,59	29,96	19,28	45,15
niepełnościenny	47	47,81	20,60	41,67	28,75	66,77
pełnościenny	81	68,09	44,05	49,40	37,25	96,43
Ogół grp.	200	50,88	35,72	41,17	26,77	64,43

Na podstawie zamieszczonych w tabeli informacji możemy między innymi zauważyć podwyższony przeciętny poziom tego parametru u osób z zawałem, zwłaszcza u tych, u których stwierdzono wystąpienie zawału pełnościennego. Znów pojawia się pytanie, czy stwierdzony stan jest efektem określonej prawidłowości?

Wreszcie ostatni z celów wymaga zbudowania modelu klasyfikacyjnego, który pozwoli ująć jednoczesny wpływ branych pod uwagę zmiennych jakościowych i ilościowych. Do zbudowania takiego modelu można zastosować wiele różnych metod (np. uogólnioną analizę dyskryminacyjną, liniowy model prawdopodobieństwa, drzewa klasyfikacyjne i inne). W naszym przykładzie zastosujemy uogólniony model analizy dyskryminacyjnej.

Poniżej przedstawiono ostateczne wyniki testowania istotności wpływu poszczególnych wprowadzonych do modelu zmiennych objaśniających.

Copyright © StatSoft Polska 2006

Efekt	Wyniki wielowymiarowych testów istotności					
	Test	**Wartość**	F	Efekt df	Błąd df	p
Wyraz wolny	Wilksa	0,58080	68,208	2	189	0,0000
LDL	Wilksa	0,73542	33,997	2	189	0,0000
Tg	Wilksa	0,96318	3,612	2	189	0,0289
PAI-1	Wilksa	0,82474	20,081	2	189	0,0000
Palenie	Wilksa	0,95958	3,981	2	189	0,0203

Wpływ wszystkich zamieszczonych w tabeli zmiennych okazał się statystycznie istotny przy poziomie prawdopodobieństwa testowego niższym od 0,05.

Jednym z powszechnie stosowanych sprawdzianów trafności modelu jest zestawienie klasyfikacji przypadków, obserwowanej w rzeczywistych danych, z klasyfikacją, jaką daje zbudowany model. Wyniki przedstawia poniższa macierz klasyfikacji przypadków.

Klasa	Macierz klasyfikacji (DaneZawały) Wiersze: obserwowana klasyfik. Kolumny: przewidywana klasyfikacja			
	Procent Poprawne	bez zawału p=,3590	niepełnościenny p=,2308	pełnościenny p=,4103
bez zawału	78,6	55	6	9
niepełnościenny	35,6	11	16	18
pełnościenny	65,0	16	12	52
Ogół	63,1	82	34	79

Jak widać, model najlepiej radzi sobie z przewidywaniem przypadków, u których nie wystąpił zawał serca (model prawidłowo klasyfikuje blisko 80% badanych osób). Nieco gorzej wygląda przewidywanie przypadków zawału pełnościennego (65% poprawnych klasyfikacji). Natomiast model zupełnie sobie nie radzi z klasyfikacją pacjentów, u których wystąpił zawał niepełnościenny. Czy można na tej podstawie twierdzić, że diagnozowanie zawału tego rodzaju jest trudniejsze? To jest problem, dla którego rozwiązania trzeba koniecznie skonsultować się ze specjalistą w zakresie kardiologii.

Zaprezentowane przykładowe wyniki analizy danych medycznych miały na celu zilustrowanie ewentualnych korzyści, jakie może odnieść badacz stosujący określone metody opracowania zebranych danych empirycznych.

Głównym jednak celem referatu było przedstawienie analizy danych jako wieloetapowego procesu, który wymaga szerszego spojrzenia na rozwiązywany problem badawczy oraz pokazanie wybranych narzędzi programu *STATISTICA* umożliwiających wspomaganie poszczególnych etapów.

Wnioski końcowe

Podsumowując przedstawione w artykule zagadnienia, można pokusić się o wyciągnięcie następujących wniosków końcowych:

- Analiza danych jest bardzo ważnym elementem składowym procesu badawczego, towarzyszącego rozwiązywaniu problemów wymagających badań empirycznych,

 Copyright © StatSoft Polska 2006

i w związku z tym nie powinna być przeprowadzana w oderwaniu od szerszego kontekstu badawczego; brak lub niewystarczająca wiedza na temat całości procesu badawczego znacznie utrudnia (a czasami wręcz uniemożliwia) przeprowadzenie analizy danych.

- Samodzielne przeprowadzenie analizy danych lub aktywne jej śledzenie wzbogaca warsztat metodologiczny badacza, zwłaszcza w kontekście projektowania przyszłych badań.
- Najtrudniejsze momenty (etapy) opracowania danych to dobór odpowiednich metod analizy (wymagający przełożenia pytań i hipotez badawczych na język analizy danych) oraz przełożenie wyników analiz na wnioski merytoryczne.
- Obecnie coraz trudniej wyobrazić sobie przeprowadzenie analizy danych bez wspomagania odpowiednimi środkami informatycznymi.
- Programy z rodziny *STATISTICA* zawierają odpowiednie narzędzia wspomagające wszystkie etapy analizy danych.

Literatura

1. Hand D., Manilla H., Smyth P., Eksploracja danych, WNT 2005.
2. Koronacki J., Mielniczuk J., Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT 2001.
3. Mackiewicz R., Francuz P., Liczby nie wiedzą, skąd pochodzą. Przewodnik po metodologii i statystyce, nie tylko dla psychologów, Wydawnictwo KUL 2005.
4. Rao C. R., Statystyka i prawda, PWN 1994.
5. Tukey J. W., Exploratory Data Analysis, Reading, MA, Addison-Wesley 1977.
6. Watała C., Biostatystyka – wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych, α-medica press, 2002.

Copyright © StatSoft Polska 2006